Webology (ISSN: 1735-188X)
Volume 18, Number 6, 2021

Comparative Analysis Of Supervised Machine Learning Algorithms For Prediction Of Orthopedic Disease

Dr.D.UmaDevi[1] , Dr.D.N.D.Harini[2] , Dr.Ch.Sita Kumari[3] , Mohammed Afeeda[4]

1 Associate Professor & Associate Head, Dept. of CSE (AI&ML), Gayatri Vidya Parishad College of Engineering (A)

2 Associate Professor & Head, Dept. of CSE, Gayatri Vidya Parishad College of Engineering (A)

3 Sr. Assistant Professor, Dept. of CSE, Gayatri Vidya Parishad College of Engineering (A)

4 M.C.A. Student, Gayatri Vidya Parishad College of Engineering (A)

Abstract-This study presents the classification of orthopedic disease patients by using the lumber and pelvic state information. Orthopedic diseases are common for people of all ages in the present day. In this study, a dataset gathered from Kaggle containing data of 310 patients with six biomechanical features describing the state of patients. Machine learning algorithms play a vital role in designing high-performance diagnosis systems and the prediction of diseases. For this purpose Logistic Regression, K-Nearest Neighbor, Random forest classifier, Decision Tree classifier algorithms are applied to the dataset. The algorithm results are then compared, the one that furnished the best result with an accuracy of 97 percent is seen as Decision Tree when compared to other algorithms' accuracy.

Keywords: Machine-learning algorithms, Orthopedic disease, Accuracy, Classification, Prediction

I. INTRODUCTION

Orthopedic diseases are caused not only due to loss of bone density but also due to deficiency of vitamin D. It occurs when the body's immune system mistakenly starts attacking its tissues and cells. It affects the musculoskeletal system which includes bones, muscles, nerves, joints, and other connective tissues. The loss of muscle tissue causes weakness and difficulty moving. This builds

the danger of cracks. Although leading technologies are being developed to diminish the pain of patients influenced by orthopedic diseases. Early counteraction is supposed to be increasingly effective in healing the condition and guaranteeing a total recovery [1].

Machine learning is a slanting methodology for early prediction of various diseases precisely dependent on the clinical parts of patients. With the information extending in the medical science field, top to bottom examination is currently possible with more precision rate and better discoveries. The emerging illness in disfigurements in bones and muscles ruins a person's development and capacity to perform everyday undertakings. Due to the lack of innovation and early preventions, many individuals endure muscle pain and, in a more regrettable case, face substantially more sicknesses. Consequently, the motivation behind this examination is to give more bits of knowledge on the prediction of orthopedic diseases ahead of schedule for counteraction of further developed diseases and help the sickness from spreading a lot in patients.

Among different orthopedic diseases, spondylolisthesis and hernia are the most frequent ones that affect individuals of all ages. A hernia happens when an organ or oily tissue presses through a weak spot in a muscle or connective tissue. It harms the lower portion of the human body and causes extreme pain. A hernia happens because of the weight of an undesirable organ into an open or frail piece of muscle or tissue. Moreover, the absence of nourishment is one of the causes of hernia. In the case of the occurrence of spondylolisthesis, the spinal bones are influenced, causing extreme agony while developing in serious cases. Spondylolisthesis is seen more in senior individuals and is gained through heredity and everyday way of life decisions. Even though mild spondylolisthesis can be cured moderately well, extreme cases require surgical treatments with the most extreme care.

In this paper, the orthopedic disease has been characterized into normal, hernia, and spondylolisthesis dependent on six features depicting the condition of the pelvic and lumber of 310 patients. The six features contain pelvic incidence, pelvic tilt, lumbar lordosis angle, sacral slope, the pelvic range also, degree spondylolisthesis. The outcomes are then trained and tested utilizing four machine learning algorithms and are assessed through the accuracy gained from the prediction models. The fundamental objective of this study is to give the best algorithm that can effectively classify and predict the diseases to help the medical specialists for early diagnosis and prevention.

The rest of the paper is composed in the accompanying request: Section II shows the literature survey behind past works done in the medical field and machine learning. Section III outlines the review and selection procedure of the algorithms and the dataset. Section IV depicts the discoveries of the prediction model; lastly, section V concludes the paper by summing up the whole study.

II. LITERATURE REVIEW

A few examinations have observed a vast contribution of machine learning in the medical field. Machine learning is utilized for an early forecast of diseases by which numerous complexities can

be prevented. For actively dealing with various levels of diseases and making the ideal choice, the most advanced way is to utilize machine learning algorithms.

One of the studies gathered dataset from an online site named Kaggle. They used Logistic Regression and other supervised machine learning algorithms for improving accuracy and the Random forest algorithm gave them the best accuracy with 89 percent [2]. Other research collected dataset from magnetic resonance images (MRI) and separated the dataset into three parts which are disk hernia, spondylolisthesis, and normal. They utilized SVM and different classifiers for improving accuracy, and the feed-forward back propagation neural network gave them the best accuracy [3]. Likewise, the perception of heart disease prediction k-Nearest Neighbor (KNN) and the genetic algorithm has been proven to be highly qualified and productive with an extra comparison between various neighbors for KNN [4]. Also, Artificial Neural Network (ANN) performed superior to Logistic regression with an accuracy of 95.8 percent in a study case for predicting the disease called Lumbar Disk Herniation (LDH) [5]. They finished up that ANN can be utilized for future decision-making procedures of medical specialists. Moreover, an artificial neural network has been utilized in research to classify patients enduring cervical disc herniation and gave excellent results over accuracy with 97 percent [6]. Similar research has been accomplished for predicting spinal fusion cost utilizing Naive Bayesian, Support Vector Machine (SVM), Logistic Regression, Random Forest, and C4.5 Decision Tree [7]. The analysis uncovered that Random Forest demonstrated to be the best classifier with high accuracy of 80 percent. All things considered, machine learning has been utilized as one of the most ingenious computational tools in the health care community.

Even though reviews have been done in the medical field with a propelled data exploration using machine learning algorithms, orthopedic disease prediction, there is as yet a moderately new zone and should be explored further for the accurate prevention and cure. Consequently, our research gives exceptional insight into the previously mentioned case and gives a comparison for the best-fitted model on account of anticipating orthopedic diseases.

III. RESEARCH METHODOLOGY

A. Data Description

The dataset utilized in this study has been gathered from the online website named Kaggle [8], containing 310 details of patients with six features to depict the shape and orientation of their pelvic and lumber gained from medical clinics. Fig. 1 shows the scatter plot of the dependent variable utilizing two biomechanical highlights that are picked haphazardly.

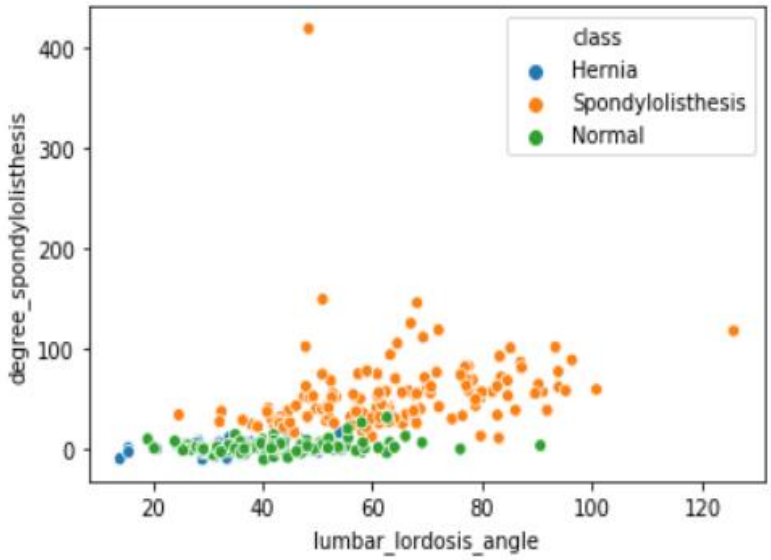

Fig.1. Scatter plot for orthopedic disease classification

The primary feature of the dataset, pelvic incidence, is an estimation of angle to decide the overall situation of the sacral plate regarding femoral heads. It is demonstrated to be a more successful measure than the pelvic base angle [9]. Pelvic tilt numeric is the second feature that depicts how much the pelvis is arranged from the thigh bones. Next, the lumber lordosis edge describes the forward low backbend angle of a human body. The sacral scope is the angular estimation of the predominant endplate and the level plane of the body. Pelvic radius is a line from the hip pivot to the back corner of the endplate. Finally, degree spondylolisthesis outlines the seriousness of spondylolisthesis of a patient. As per these six features, the data populace has been classified into normal, hernia, and spondylolisthesis. The disease class has been resolved and four machine learning algorithms, in particular Logistic Regression, k-Nearest Neighbor, Random Forest classifier, and Decision Tree classifier, were run on the relating dataset for accurate prediction. The dataset had no invalid or missing values and contained no exceptions or qualities that surpassed the input range. The dataset was split into two sections: train part and test part. Eighty percent of the dataset was considered as the training set and the staying twenty percent was considered as a test set.

B. Algorithm Implementation

Since the input and output features were at that point present in the dataset, it required a supervised machine learning technique to deal with the mapping function. The dataset was separated into 80 percent for training and 20 percent for testing since a more prominent bit of data for the training set gives better classification and prediction models. All the algorithms are controlled by utilizing Spyder IDE in Anaconda application and were reevaluated by WEKA 3.8.1 data mining software.

Logistic Regression: This strategy is utilized for both classification and regression. It is mostly utilized because it overcomes the limitation of linear regression by creating results for

a linear relationship yet additionally for continuous data patterns. This algorithm is a supervised approach and can be utilized to predict unmitigated results. This algorithm expects to find the ideal fitting model to predict outcome variables using a set of dependent variables. Fig.2 shows Pseudo-code for logistic regression.

1: Start with random weights: $w_1, \ldots, w_n, b$
2: **for** every point $(x_1, x_2, \ldots, x_n)$: **do**
3: **for** $i = 1, 2, \ldots, n$ **do**
4: Update $w_i^{'} \leftarrow w_i - \alpha(\hat{y} - y)x_i$
5: Update $b^{'} \leftarrow b - \alpha(\hat{y} - y)$
6: Repeat until error is small

Fig.2. Pseudo-code for logistic regression

K-Nearest Neighbor: This is a non-parametric lazy algorithm since it has no pre-suspicion about the dataset. It takes the most widely recognized neighbor among a total of k-neighbors and relegates it to the classification. The accuracy relies upon the neighbor value (k). This algorithm is considered extraordinary compared to other classification algorithms [10]. Fig. 3 shows the pseudo-code for K-Nearest Neighbor (KNN).

k-Nearest Neighbor
Classify (X, Y, *x*) // X: training data, Y: class labels of X, *x*: unknown sample
for *i* = 1 **to** *m* **do**
Compute distance $d(\mathbf{X}_i, x)$
end for
Compute set *I* containing indices for the *k* smallest distances $d(\mathbf{X}_i, x)$.
return majority label for $\{\mathbf{Y}_i$ where $i \in I\}$

Fig.3. Pseudo-code for k-Nearest Neighbor

Random Forest Tree: This algorithm works by averaging various decision trees at training time. It delivers a more accurate result than the decision tree since it picks up the outcome from various example cases. The number of trees created in this classifier is determined to 150 by the parameter n_estimator, and the max_depth parameter is set as 6, which is the most extreme profundity of nodes for the trees. Fig. 4 shows the pseudo-code for Random Forest Classifier.

Algorithm 1 Random Forest

Precondition: A training set $S := (x_1, y_1), \ldots, (x_n, y_n)$, features F, and number of trees in forest B.

```
function RandomForest(S, F)
    H ← ∅
    for i ∈ 1, ..., B do
        S^(i) ← A bootstrap sample from S
        h_i ← RandomizedTreeLearn(S^(i), F)
        H ← H ∪ {h_i}
    end for
    return H
end function
function RandomizedTreeLearn(S, F)
    At each node:
        f ← very small subset of F
        Split on best feature in f
    return The learned tree
end function
```

Fig.4. Pseudo-code for Random forest classifier

Decision Tree: Decision Trees are a non-parametric supervised learning method utilized for both classification and regression tasks. It is one of the most generally used and practical techniques for supervised learning. Decision trees are developed through an algorithmic methodology that recognizes approaches to split a data set dependent on various conditions. Fig.5 shows the pseudo-code for decision tree classifier.

```
INPUT: S, where S = set of classified instances
OUTPUT: Decision Tree
Require: S ≠ ∅, num_attributes > 0
 1: procedure BuildTree
 2:     repeat
 3:         maxGain ← 0
 4:         splitA ← null
 5:         e ← Entropy(Attributes)
 6:         for all Attributes a in S do
 7:             gain ← InformationGain(a, e)
 8:             if gain > maxGain then
 9:                 maxGain ← gain
10:                 splitA ← a
11:             end if
12:         end for
13:         Partition(S, splitA)
14:     until all partitions processed
15: end procedure
```

Fig.5. Pseudo-code for decision tree classifier

For analyzing the accuracy of the prediction, confusion matrix and prediction accuracy in percentage have been deciphered. Confusion Matrix is a matrix portrayal of the achievement of the algorithm. In a two-by-two dimension, confusion matrix can be depicted by four

parameters which are true positives, true negatives, false positives, and false negatives. True positives are the samples that we predicted as true and their real qualities are likewise evident. True negatives are the samples which we predicted as false and their actual values are likewise false which implies they don't have the sickness in this situation. False positives are the samples which we predicted as true yet they don't have the disease and false negatives are the samples which we predicted as false but they have the disease. Likewise, accuracy is a rate proportion of performance of the classification model which has been utilized in this study. Fig.6 shows the pictorial representation of the concept of the confusion matrix.

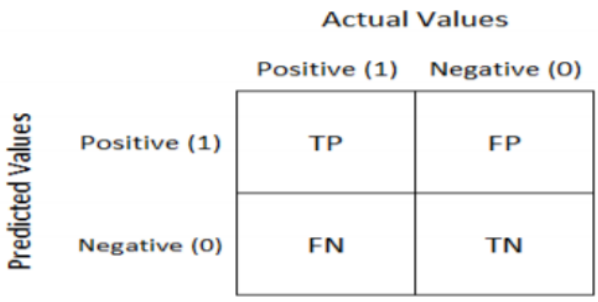

Fig.6. Pictorial representation of confusion matrix

IV. RESULT ANALYSIS

For various neighbor values of the k-Nearest-Neighbor algorithm (KNN), the best accuracy is picked up for a K-estimation of 3. Table I represents the accuracy comparison between six different k-values where k-value means the neighbor value used in the k-Nearest Neighbor algorithm. Since the complete classification for the orthopedic disease is 3, the most elevated accuracy is picked up when the data is divided into 3 neighbors and the nearest neighbor is picked for decision making.

k-value	Accuracy (%)
1	74
2	85
3	86
4	83
5	76
6	77

TABLE I. DIFFERENCE INACCURACY FOR NEIGHBOR VALUES

After running the four previously mentioned algorithms, the accompanying results are picked up. Accuracy is the performance measure of various machine learning algorithms. In this study, the accuracy of the various algorithms marginally contrasts from one another as their working mechanism is not quite the same as one another. Table II represents the accuracy of predicting the referenced disease of the model.

Algorithms	Accuracy (%)
Logistic Regression	82
k-Nearest Neighbor(k=3)	84
Random Forest	93
Decision Tree	97

TABLE II. ACCURACY OF ALGORITHMS

The accuracy rate gives us how effectively the prediction model is performing regarding the actual outcome. Fig.7 shows the accuracy comparison of the four algorithms.

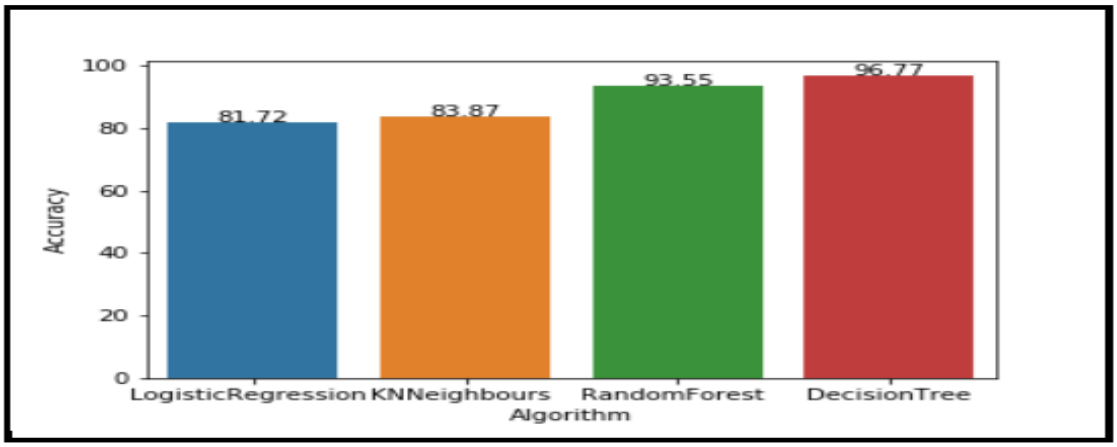

Fig.7. Accuracy comparison of four algorithms

As we fixed the test set to 20 percent out of the 310 samples, the total sample for the test is 62. The perceptions are that for logistic regression, the system can predict 50 results effectively out of 62 estimations of the test set accurately, and using k=3 in the k-Nearest Neighbor algorithm, 53 results can be predicted correctly out of 62 samples. Random Forest can predict 55 results efficiently out of 62 values and decision trees can predict 58 outcomes out of 62 samples. Table III represents the confusion matrix of each of the previously mentioned algorithms where N represents Normal, H represents Hernia and S represents Spondylolisthesis.

Algorithm	Confusion Matrix
Logistic Regression	N H S N 8 3 0 H 5 14 0 S 0 1 31
k-Nearest Neighbor(k=3)	N H S N 9 2 0 H 6 13 0 S 1 1 30

Random Forest	N H S N 9 2 0 H 3 16 0 S 0 1 31
Decision Tree	N H S N 7 1 0 H 3 11 0 S 0 1 30

TABLE III. CONFUSION MATRIX OF ALGORITHMS

From the above tables, the Decision Tree classifier produces the best accuracy of 97 percent and gives an excellent confusion matrix among all the algorithms for this dataset.

V. CONCLUSION

Orthopedic diseases render the everyday lives of the individuals by making them incapable to process day by day exercises as a typical individual would. Early diagnosis and treatment may prevent the furthest point of the orthopedic disease and give a better fix to the patients. By using machine learning approaches, it is conceivable to do as such, and analysts are constantly attempting to better their results. This study intends to give more experiences on the matter and gives that among Logistic Regression, K Nearest Neighbor, Random Forest, and Decision Tree, a Comparison of these four models clearly express that Decision Tree gives the best accuracy of 97 percent and can be used for the clinical fields.

Although the accuracy of higher than 80 percent has been gained for every one of the models, the dataset is as yet constrained and requires further processing and a larger population to draw a much stronger conclusion. The future examination includes assembling more data and exploring more algorithms to give a point by point distinguishing proof and implementation of the prediction model.

REFERENCES

[1] Blatter and J. Dvorak, "Football for health - prevention is better than cure", Scandinavian Journal of Medicine & Science in Sports, vol. 20, p. v-v, 2010.

[2] Nadia Rubaiyat, Anika Islam Apsara, Abdullah Al Farabe and Ifaz Isthiak, "Classification and prediction of Orthopedic disease based on lumber and pelvic state of patients", 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT).

[3] P. Srimani and M. Koti, "Medical Diagnosis Using Ensemble Classifiers - A Novel Machine-Learning Approach", Journal of Advanced Computing, 2013.

[4] M. Jabbar, B. Deekshatulu, and P. Chandra, "Classification of Heart Disease Using K-Nearest Neighbor and Genetic Algorithm", Procedia Technology, vol. 10, pp. 85-94, 2013.

[5] P. Azimi, E. Benzel, S. Shahzadi, S. Azhari, and A. Zali, "Prediction of Successful Surgery Outcome in Lumbar Disc Herniation Based on Artificial Neural Networks", Global Spine Journal, vol. 4, no. 1, pp. s-0034-1376643-s-0034-1376643, 2014.

[6] G. Ozmen, "Classification of Cervical Disc Herniation Disease using Muscle Fatigue based surface EMG signals by Artificial Neural Networks", International Journal of Intelligent Systems and Applications in Engineering, vol. 4, no. 5, pp. 256-262, 2017.

[7] F. Cabitza, A. Locoro and G. Banfi, "Machine Learning in Orthopedics: A Literature Review", Frontiers in Bioengineering and Biotechnology, vol. 6, 2018.

[8] Dataset Available:https://www.kaggle.com/uciml/biomechanical-features-of-orthopedic-patients

[9] S. Ramchandran, "Pelvic Incidence (PI) is more Easily Understood as the Pelvic Base Angle (PBA)", Spine Research, vol. 03, no. 01, 2017.

[10] G. Chen and D. Shah, "Explaining the Success of Nearest Neighbor Methods in Prediction", Foundations and Trends® in Machine Learning, vol. 10, no. 5-6, pp. 337-588, 2018.

[11]C. Kuo, L. Yu, H. Chen, and C. Chan, "Comparison of Models for the Prediction of Medical Costs of Spinal Fusion in Taiwan Diagnosis Related Groups by Machine Learning Algorithms", Healthcare Informatics Research, vol. 24, no. 1, p. 29, 2018.